

CLAIMS

1. A method of determining parameter combinations for automated access to World Wide Web content that is accessible based on parameters resulting from real user interactions with a World Wide Web site, said method comprising:

maintaining at least one log file containing at least one set of parameters resulting from real user interactions with said World Wide Web site;

analyzing said log file to determine parameter combinations for automated access to said World Wide Web content.

2. A method of determining parameter combinations for automated access to World Wide Web content that is accessible based on parameters resulting from real user interactions with a World Wide Web site, as per claim 1, wherein said parameters are entries in HTML forms, said analyzing step further comprising

ranking entries in each set of entries according to their frequency of occurrence;

for each set of entries resulting from unlimited text entries, excluding entries ranked below a predetermined number; and

wherein said parameter combinations are determined by producing combinations of entries from each set of entries.

3. A method of determining parameter combinations for automated access to World Wide Web content that is accessible based on parameters resulting from real user interactions with a

3 World Wide Web site, as per claim 2, wherein said parameter combinations are determined by
4 producing all combinations of entries from each set of entries.

1 4. A method of determining parameter combinations for automated access to World Wide
2 Web content that is accessible based on parameters resulting from real user interactions with a
3 World Wide Web site, as per claim 2, wherein entries resulting from limited text entries and
4 unlimited text entries have stop words removed and remaining words stemmed.

5 5. A method of determining parameter combinations for automated access to World Wide
6 Web content that is accessible based on parameters resulting from real user interactions with a
7 World Wide Web site, as per claim 1, wherein said log file is maintained by a proxy server that
8 logs communications between a client and a Web server resulting from real user accesses to said
9 World Wide Web content.

10 6. A method of determining parameter combinations for automated access to World Wide
11 Web content that is accessible based on parameters resulting from real user interactions with a
12 World Wide Web site, as per claim 1, wherein said content is automatically accessed using said
13 parameter combinations.

1 7. A method of increasing web crawler penetration of Web databases accessible via HTML
2 forms, said method comprising:
3 reviewing previous real user queries;

identifying possible queries for said Web crawler from said previous real user queries by synthesis of entries for any of: predefined sets, limited text entries or unlimited text entries; and

providing said identified queries to said Web crawler during an instantiation of automated access to said Web databases by said Web crawler.

8. A method of increasing web crawler penetration of Web databases accessible via HTML forms, as per claim 7, wherein said previous user queries are maintained in a log file.

9. A method of increasing web crawler penetration of Web databases accessible via HTML forms, as per claim 8, wherein said log file is maintained by a proxy server.

10. A method of increasing web crawler penetration of Web databases accessible via HTML forms, as per claim 7, wherein said synthesis comprises:

ranking any entries for predetermined sets;

ranking any entries for limited text entries;

ranking any entries for unlimited text entries;

excluding entries for unlimited text entries ranked below a predetermined number;

and

pairing entries from each set of ranked entries.

11. A method of increasing web crawler penetration of Web databases accessible via HTML forms, as per claim 10, wherein said synthesis further comprises:

3 removing stop words and stemming remaining words for entries resulting from
4 limited text entries and unlimited text entries.

1 12. A method of determining entries for input items of an HTML form for automated
2 accesses to content contained in a Web database behind said HTML form, said method
3 comprising:

4 maintaining a log of real user entries for said input items;

5 analyzing said log to determine entry combinations for said input items.

6 13. A method of determining entries for input items of an HTML form for automated
7 accesses to content contained in a Web database behind said HTML form, as per claim 12,
8 wherein said log file contains at least one set of entries, said analyzing step further comprising
9 ranking entries in each set of entries according to their frequency of occurrence;
10 for each set of entries resulting from unlimited text entries, excluding entries
11 ranked below a predetermined number; and
12 wherein said automated parameter combinations are determined by producing
13 combinations of entries from each set of entries.

1 14. A method of determining entries for input items of an HTML form for automated
2 accesses to content contained in a Web database behind said HTML form, as per claim 13,
3 wherein said parameter combinations are determined by producing all combinations of entries
4 from each set of entries.

1 15. A method of determining entries for input items of an HTML form for automated
2 accesses to content contained in a Web database behind said HTML form, as per claim 13,
3 wherein entries resulting from limited text entries and unlimited text entries have stop words
4 removed and remaining words stemmed.

1 16. A method of determining entries for input items of an HTML form for automated
2 accesses to content contained in a Web database behind said HTML form, as per claim 12,
3 wherein said log file is maintained by a proxy server that logs communications between a client
4 and a Web server resulting from real user accesses to said World Wide Web content.

1 17. A method of emulating real user access to World Wide Web content dynamically
2 accessible via an HTML form, said method comprising:

3 maintaining a log containing real user entries into each input item of said HTML
4 form;

5 ranking entries for each input item according to their frequency of occurrence;

6 for each unlimited text entry input item, excluding entries ranked below a
7 predetermined number;

8 determining combinations of entries from each set of entries; and

9 automatically accessing said content using said combinations of entries.

1 18. A method of emulating real user access to World Wide Web content dynamically
2 accessible via an HTML form, as per claim 17, wherein entries resulting from limited text entries
3 and unlimited text entries have stop words removed and remaining words stemmed.

1 19. A method of emulating real user access to World Wide Web content dynamically
2 accessible via an HTML form, as per claim 17, wherein said log file is maintained by a proxy
3 server that logs communications between a client and a Web server resulting from real user
4 accesses to said World Wide Web content.

1 20. An article of manufacture comprising a computer usable medium having computer
2 readable program code embed therein to determine parameter combinations for automated access
3 to World Wide Web content that is accessible based on parameters resulting from user
4 interactions with a World Wide Web site, said computer readable program code comprising:

5 computer readable program code for maintaining at least one log file
6 representative of real user interactions with said World Wide Web site;

7 computer readable program code for analyzing said log file to determine
8 parameter combinations for automated access to said World Wide Web content.

1 21. An article of manufacture comprising a computer usable medium having computer
2 readable program code embed therein to determine parameter combinations for automated access
3 to World Wide Web content that is accessible based on parameters resulting from user
4 interactions with a World Wide Web site, as per claim 20, wherein said parameters are entries in
5 HTML forms, said computer readable program code for analyzing further comprising

6 computer readable program code for ranking entries in each set of entries
7 according to their frequency of occurrence; and

8 computer readable program code for each set of entries resulting from unlimited
9 text entries, excluding entries ranked below a predetermined number; and
10 wherein said parameter combinations are determined by producing combinations
11 of entries from each set of entries.

1 22. An article of manufacture comprising a computer usable medium having computer
2 readable program code embed therein to determine parameter combinations for automated access
3 to World Wide Web content that is accessible based on parameters resulting from user
4 interactions with a World Wide Web site, as per claim 21, wherein said parameter combinations
5 are determined by producing all combinations of entries from each set of entries.

6 23. An article of manufacture comprising a computer usable medium having computer
7 readable program code embed therein to determine parameter combinations for automated access
8 to World Wide Web content that is accessible based on parameters resulting from user
9 interactions with a World Wide Web site, as per claim 21, wherein entries resulting from limited
4 text entries and unlimited text entries have stop words removed and remaining words stemmed.

1 24. An article of manufacture comprising a computer usable medium having computer
2 readable program code embed therein to determine parameter combinations for automated access
3 to World Wide Web content that is accessible based on parameters resulting from user
4 interactions with a World Wide Web site, as per claim 20, wherein said log file is maintained by
5 a proxy server that logs communications between a client and a Web server resulting from real
6 user accesses to said World Wide Web content.

1 25. An article of manufacture comprising a computer usable medium having computer
2 readable program code embed therein to determine parameter combinations for automated access
3 to World Wide Web content that is accessible based on parameters resulting from user
4 interactions with a World Wide Web site, as per claim 20, wherein said content is automatically
5 access using said parameter combinations.